A nonparametric approach for synthetic multisite streamflow generation

JG Ndiritu

School of civil and Environmental Engineering, University of the Witwatersrand, P Bag 3, WITS, 2050, South Africa

Abstract

A nonparametric method for synthetic multisite streamflow generation is developed, tested and compared with the STOMSA time series-based model. The nonparametric model and STOMSA both reproduce the statistics of historical flows reasonably although the nonparametric method has the tendency to slightly overestimate the standard deviation of annual flows. STOMSA fails to reproduce the monthly serial correlation between the end of one year and the beginning of the next for all five sites tested while the nonparametric method models this correlation successfully. STOMSA is also found to underestimate the annual serial correlation while no bias is detected with the nonparametric method. Using yield - storage capacity test, the nonparametric model is found to be satisfactory but may have the tendency to overestimate capacities for large yields. STOMSA gives reasonably unbiased capacities results for all five sites. The minimum cumulative flows test for durations varying from 1 to 24 years, gives similar results with the nonparametric model and STOMSA. The nonparametric method however gives a wider range of cumulative flows.

Keywords*: nonparametric, synthetic streamflow generation, multisite*

1. Introduction

Stochastic sequences are invaluable for evaluating the reliability of water resource systems. Historical streamflows provide a guide as to the flows that will be expected during the life of the system but will hardly ever be repeated. By generating several synthetic sequences that are statistically similar to the historical ones, a probabilistic analysis of the system performance can be carried out to aid decisions regarding system sizing and operation. Over the decades synthetic streamflow generation methods based on time series analysis have been predominant. These methods have worked well in some regions particularly in South Africa where a robust method currently versioned as STOMSA (**STO**chastic **M**odel of **S**outh **A**frica) is used extensively in long-term planning and design and short-term operation of reservoir systems (Pegram and McKenzie, 1991, Basson et al., 1994, Van Rooyen and McKenzie, 2004). In the USA, Vogel and Shallcross (1996) indicate that the Bureau of Reclamation of the US Department of the Interior often uses a simpler nonparametric method, the index sequential method (ISM) instead of their own parametric model (LIST). The reason for this is the ease of application of the ISM method. Time series methods require one to normalize the flows. The choice of the distribution that will do this best may be problematic. Time series models also require the determination of a large number of parameters and some features such as bimodality of monthly distributions and jumps of annual flows in the historical sequences may be difficult to reproduce. Fitting of distributions to streamflows having zero or very low flows in some years is also difficult. The disaggregation methods used in many time series methods leads to the nonconservation of the dependence between the seasonal or monthly flow at the end of one year and the beginning of the next (Sharma and O'Neill, 2002). Research into the applicability of nonparametric methods for synthetic streamflow generation and their use as possible alternatives to time series methods has thus been gaining momentum (Lall and Sharma, 1996, Vogel and Shallcross, 1996, Sharma et al., 1997, Tarboton et al., 1998, Srinivas and Srinivasan, 2000, 2001, Sharma and O'Neill, 2002). In these studies, the nonparametric methods have been found to perform as well and sometimes better than parametric methods. While some of the methods are easy to understand and apply (e.g. the moving block bootstrap (Vogel and Shallcross, 1996), and to a lesser extent the nearest neighbour method (Lall and Sharma, 1996), others are complex and computation intensive (e.g. Sharma and O'Neill, 2002) and do not hold any advantage over parametric methods in these respects. The moving block bootstrap and nearest neighbour method resample from the historical data only meaning the synthetic flows will not take values beyond the range of the historical ones - an obviously significant limitation. The above mentioned studies on nonparametric methods were confined to single sites while practical water resources systems analysis often involves multiple sites thereby necessitating the preservation of cross correlations among them.

The study reported here aimed at developing and testing a nonparametric method that is simple to understand and apply and:

1. maintains the statistics of the historical sequences,
2. is capable of obtaining flows other than those in the historical sequence,
3. maintains monthly serial correlation within years and between the end of one year and the beginning of the next,
4. maintains lag 1 annual serial correlations, and
5. maintains monthly and annual cross correlations of multiple sites.

The method was tested by the simultaneous generation of synthetic sequences at five sites in South Africa and was also compared with STOMSA (**STO**chastic **M**odel of **S**outh **A**frica) (Van Rooyen and McKenzie, 2004).

2. Model Description

2.1 Derivation of annual flows

Instead of using a constant moving block length as in the moving block bootstrap method (Vogel and Shallcross, 1996), the following method that obtains variable lengths is used.

Step	Action
1	Specify a minimum block length *ml* years
2	Start at year 1 of the sequence
3	Skip *ml* years from year 1
4	Skip *ml* years from the end of the last block
5	Check the ratio of annual flow to mean flow for the next year
7	if this ratio < p_1 then move to the next year-go to step 5
6	if this ratio ≥ p_1 then
8	check the ratio of annual flow to mean flow for the next year
9	If this ratio ≥ p_2 this is the end of the block
	go to step 4 to obtain the next block
10	if this ratio < p_2 then check the ratio of annual flow to mean flow in the previous year
11	If this ratio ≥ p_3 then this is the end of the block
	go to step 4 to obtain the next block
12	If this ratio <p_3, then move to the next year – go to step 5

The method ensures that the length of each block equals or exceeds a specified number of years denoted as *ml*. The occurrence of the end of a block in the middle of a critical period is discouraged and is encouraged to happen when flows are high. This is done using the three parameters p_1, p_2 and p_3 which are ratios of the annual flow in a given year to the mean flow of the sequence. A ratio less than p_1 implies a dry year. If the ratio exceeds p_1 and that in the succeeding year exceeds a higher ratio p_2, the drought period is considered broken and the block is then terminated. If the flow in a given year is very high (with a ratio ≥ p_3), the block is terminated irrespective of the flow in the succeeding year. For the method to work, $p_1 < p_2 < p_3$.

To obtain the annual flows of a synthetic sequence, the blocks are selected randomly with replacement until the required length of the synthetic sequence is obtained.

2.2 Modelling annual cross correlations

Since the method is meant for multiple sites, a means of ensuring the maintenance of annual cross correlations is required. An attempt was made to do this by obtaining the synthetic annual flows for each sequence independently and modelling the regional droughts in the historical record using variable temporal translations of the synthetic sequences. This improved correlations but they were still considerably lower than the historical ones. It was therefore decided to use each historical sequence as the lead sequence. Firstly, the lead sequence is used to obtain a synthetic sequence of annual flows. The synthetic sequences for the other sites are then obtained using an identical sequencing of years as the synthetic sequence obtained from the lead sequence. Each historical sequence is used as the lead sequence an equal number of times.

2.3 Dissagregation of annual flows into monthly flows

The steps are described with explanations of the reasoning behind each.

1) Rank the flows of the synthetic and the historical sequence in order of magnitude. Assign a class to each flow in the sequence based on its rank. Each class is assigned an equal number of flows and the maximum number of classes used is a specified model parameter nc_{max}.
2) Pair up consecutive years in the synthetic sequence so that each year except the first and the last has two pairs, one with the previous year and one with the succeeding year.
3) For all pairs of consecutive years of the synthetic sequence, obtain a corresponding pair of consecutive years in the historical sequence. Ensure that the two pairs are identical with regards to the classification done in (1) but the flows themselves are not identical. Commence the search at a randomly selected location of the historical sequence. If no pair is obtained after running through the complete historical sequence, reduce the number of classes by 1, redo the ranking for both the synthetic and the historical sequence and repeat the search. With the completion of this selection, each year except the first and the last has two matching historical years; one from the pair obtained with the previous year (termed as the first historical match) and one with the pair obtained with the succeeding year (the second historical match). The first and the last year have one matching historical year.
4) For each year except the first and the last, obtain the ratios of the monthly to yearly flow for the two matched historical years. Compute the monthly to yearly flow ratio for each month of the synthetic sequence as a weighted value of the corresponding ratios of the two matching years. For month *i*, the weights are (12-*i*)/12 and *i*/12 for the first and the second historical match respectively as illustrated in Figure 1. With this weighting, the monthly to annual flow ratios at the end of the year match the historical ones exactly and reduce gradually towards the middle of the year as those of the neighbouring pairs are introduced at increasing weights. This helps to maintain the monthly serial correlations at end of the year. The weighting and the selection of matching years whose annual flows are not equal to the synthetic flows ensures that the computed monthly flows of the synthetic sequence are not identical to those in the historical sequence.

Moreover, when these are summed up to obtain a new annual synthetic flow, the flow is almost always likely to be different from the historical one. For the first and the last year, a monthly to annual flow ratio of the single matching years is used.

2.4 Modelling monthly cross correlations

In order to maintain monthly cross correlations, a similar approach to that in step 2.2 was found necessary. The historical matching pairs of years obtained in step 2.3 for the leading sequence are used for the other sequences.

2.5 Adjusting synthetic correlations to match historical ones

It was observed, that using the monthly to yearly flow ratios of two years to obtain the ratios for one year in step 2.3 leads to higher monthly serial correlations in the synthetic sequences than in the historical ones. This effect was found to be carried on to monthly cross correlations as well leading to higher correlations than for the historical sequences. To deal with this, perturbations are introduced to the synthetic monthly to yearly flow ratios obtained in step 2.3 until the overall monthly serial and cross correlations for all sites match the historical ones as closely as possible. The closeness is quantified by summation of the square of differences. The perturbation applied has the form ±0.5(perturbation+0.05rand(0,1)) with the perturbation varying from 0 to 0.95 and rand(0,1) being a random number between 0 and 1. This equation was obtained after a series of trials with several forms of perturbations including a more obvious form: ±perturbation(rand(0,1)-0.5). Figure 2 illustrates how the optimal perturbation is obtained.

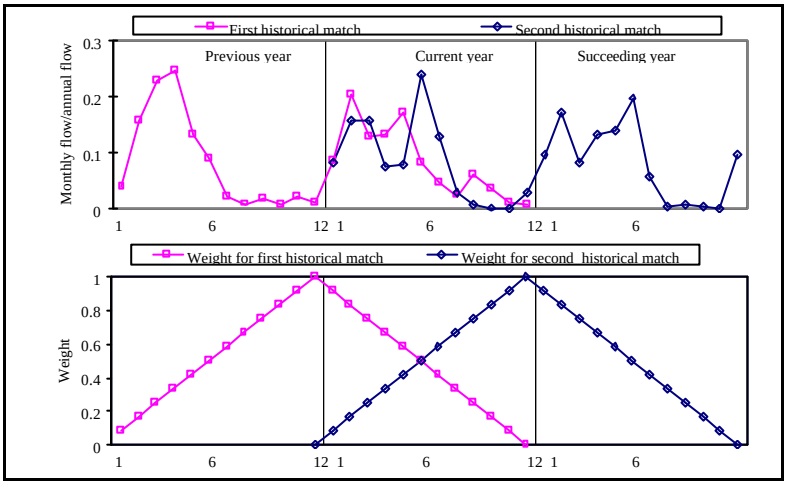


Figure 1 Illustration of the dissagregation of annual flows to monthly flows

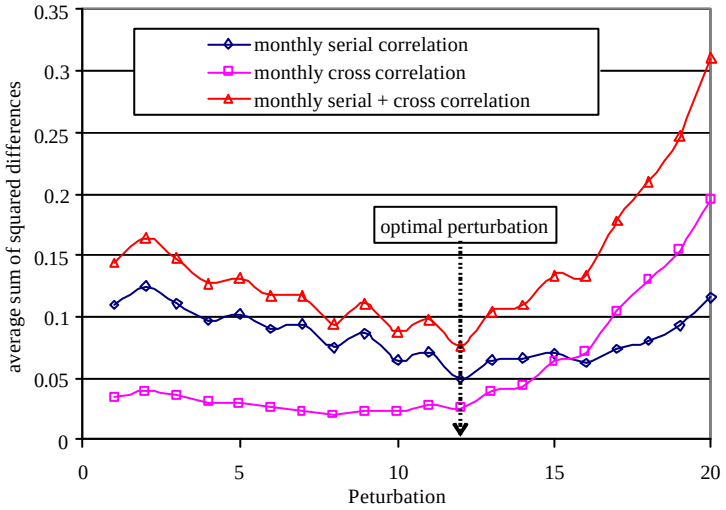


Figure 2 Matching synthetic and historical monthly correlations

3. Selection of model parameters

The five model parameters are *ml*, p_1, p_2, p_3 and nc_{max}. To enable varying streamflows to be obtained, a range within which to randomly obtain each of these parameters is specified. The generation of each sequence therefore uses a unique parameter set obtained randomly within the ranges. No comprehensive tests were carried out to assess the effects of varying the parameter ranges of these parameters and a trial and error method was applied to obtain the adopted ranges. Table presents these ranges. Note that *ml* and nc_{max} are integers while p_1, p_2 and p_3 are continuous variables.

Table 1 Applied model parameter ranges

Parameter	***ml***	p_1	p_2	p_3	nc_{max}
Range	2 - 4	0.1 - 0.5	0.6 - 1.5	2.0 – 5.0	4 - 9

4. Choice of Data

The STOMSA (**STO**chastic **M**odel of **S**outh **A**frica) guidelines (Van Rooyen and McKenzie, 2004) contain a dataset of streamflows for five incremental catchments whose characteristics are reproduced here in Table 2. This dataset was found appropriate for the task at hand and was adopted for use.

Table 2 Runoff characteristics for selected incremental sub-catchments (from Van Rooyen and McKenzie, 2004)

Description	**Start year (hydrological)**	**End year (hydrological)**	**Record period length (years)**	**Mean annual runoff (million m^3)**
Bloemhof Dam	1920	1994	75	154
Delangesdrift Dam	1920	1994	75	249
Katse Dam	1920	1995	76	546
Vaal Dam	1920	1994	75	519
Welbedacht Dam	1920	1987	68	630

5. Tests and comparison of the nonparametric method (NP) with STOMSA

The testing consisted of a verification and a validation phase as used in several other studies (e.g. Pegram and McKenzie, 1991). Verification compares the distribution of the statistics of the generated sequences with the historical statistics and validation uses tests not applied directly in setting up the model. The tests here were based on the generation of 100 sequences each 76 years long.

5.1 Verification tests

The verification results are presented graphically in Figures 3 to 8 and the notable observations are summarized in Table 3.

Table 3 Main observations from verification tests

Figure	**Observation**
3	The two methods do not give any bias in the mean of annual flows. NP has a tendency to obtain a higher range of means than STOMSA.
4	NP has the tendency to slightly overestimate the standard deviation of annual flows while STOMSA does not show any bias.
5 and 6*	No notable bias is found in the monthly means and standard deviation with both NP and STOMSA.
7	Both NP and STOMSA reproduce the annual serial correlations adequately. NP obtains a closer match with the historical values than STOMSA.
8	STOMSA is not able to reproduce the monthly serial correlation between the end of one year and the beginning of the next. The rest of the correlations are reproduced adequately by the two methods.
9	STOMSA has the tendency to underestimate the annual serial correlations while NP does not show any bias.

* - this observation is based on box plots for the other sites as well.

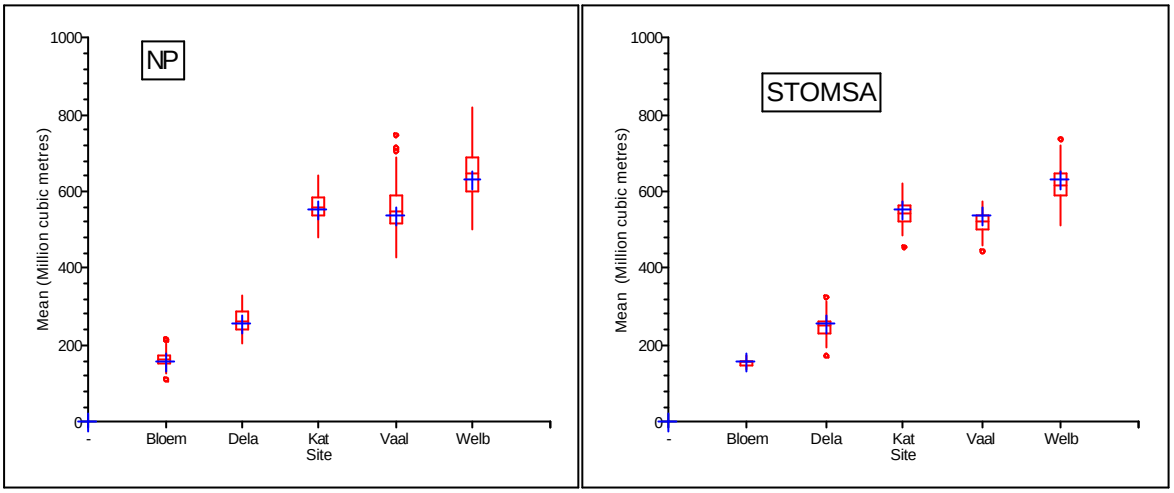


Figure 3 Box plots of annual mean flow

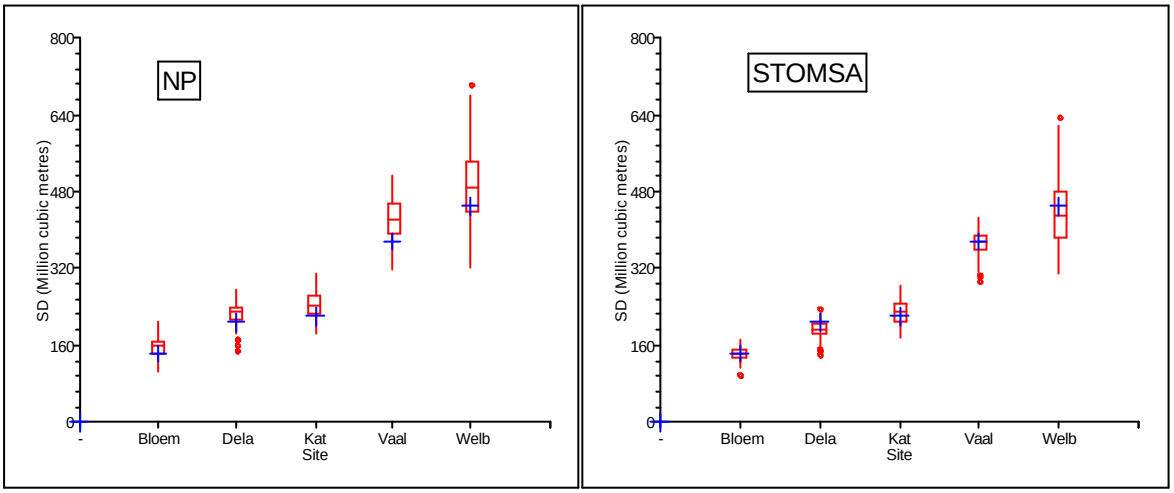


Figure 4 Box plots of annual standard deviations

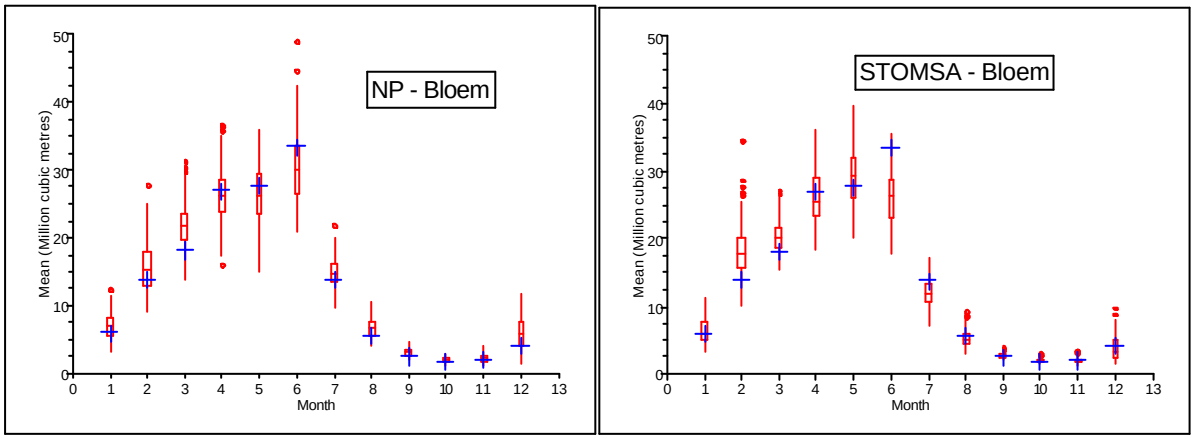


Figure 5 Box plots of monthly mean flow for Bloemhof Dam

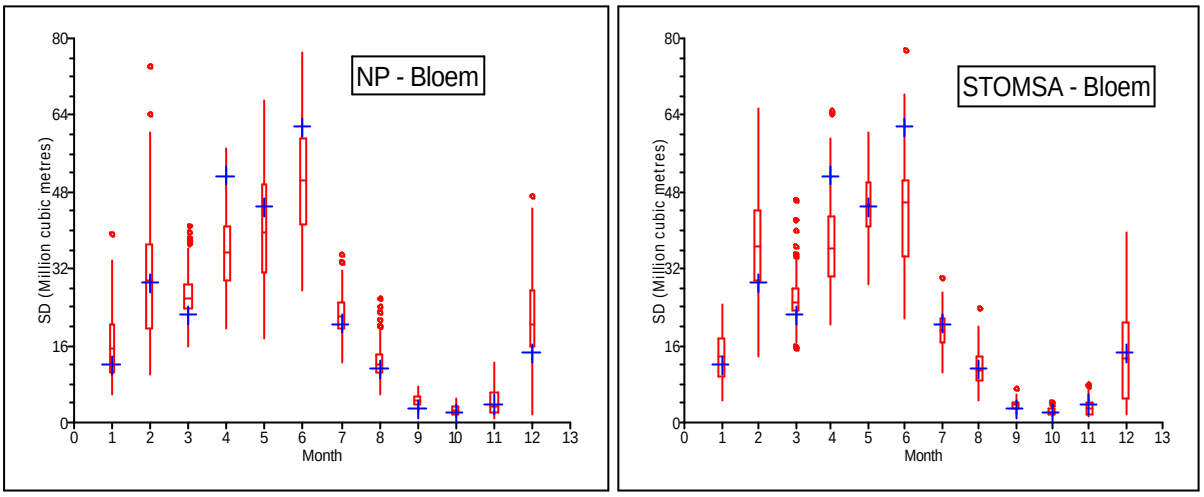
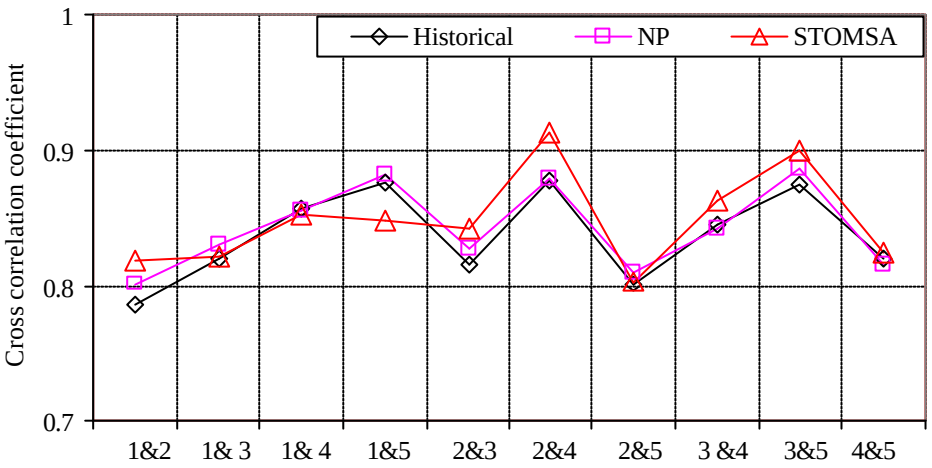


Figure 6 Box plots of monthly standard deviation for Bloemhof Dam



Pair of sequences (1-Bloem, 2-Dela, 3-Kat, 4-Vaal, 5-Welb)

Figure 7 Comparison of annual cross correlations

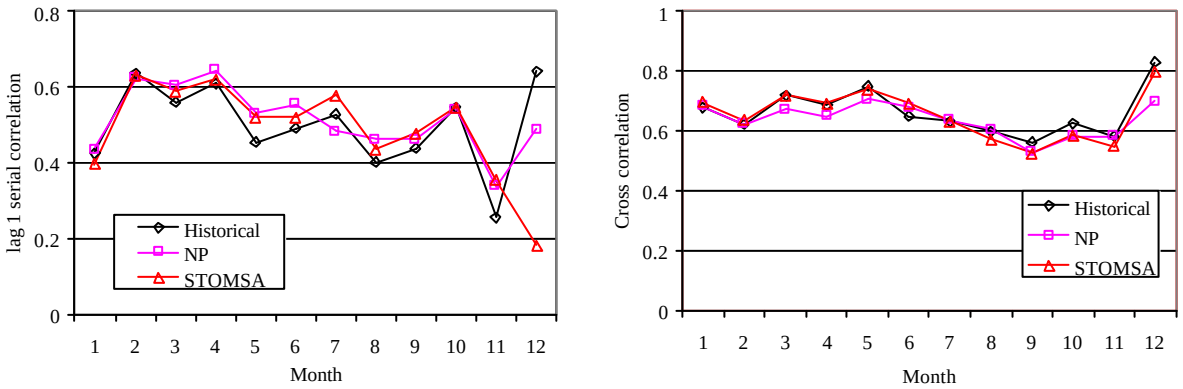


Figure 8 Comparison of overall average monthly serial and cross correlations.

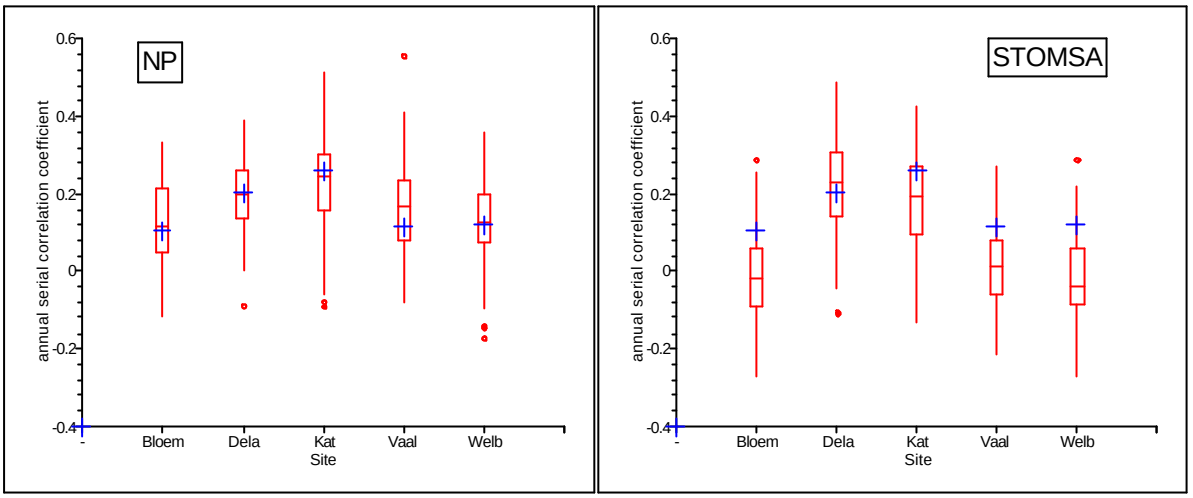


Figure 4 Box plots of annual serial correlation coefficient

5.2 Validation tests

Two tests were carried out: 1) the yield-storage capacity test using the sequent peak algorithm method for yields varying from 5 to 100% the mean annual flow (MAR) and 2) the minimum cummulative streamflow test for durations of 1 to 24 years. Some of the box plots from these tests are presented in Figures 10 to 13. The main observations from these and other box plots not presented here are summarised in Table 4.

Table 4 Main observations from validation tests

Figures	Observation
10 and 11*	NP overestimated the capacities for Bloemhof dam. The box plots of Figure 9 show that both NP and STOMSA produced values biased to the higher side to a yield of 55% but STOMSA obtained unbiased values for higher yields while NP maintained the bias. NP was also found to overestimate capacities for Katse dam for 90-100% yield. The rest of the box plots were essentially similar for the two methods.
12 and 13*	Both NP and STOMSA obtained reasonable minimum cummulative flows with NP giving wider ranges than STOMSA for 2 sites.

* The observations are based also on box plots of the other sites

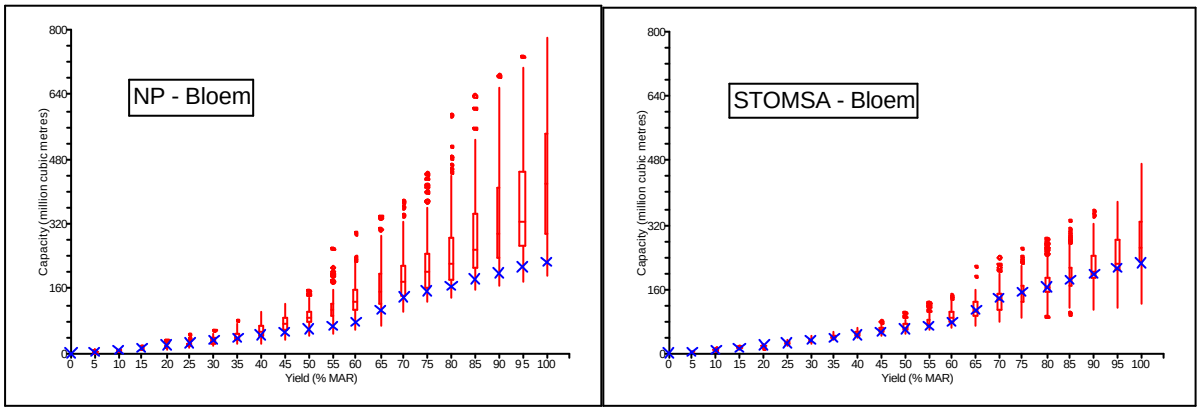


Figure 10 Box plots of capacities for Bloemhof dam

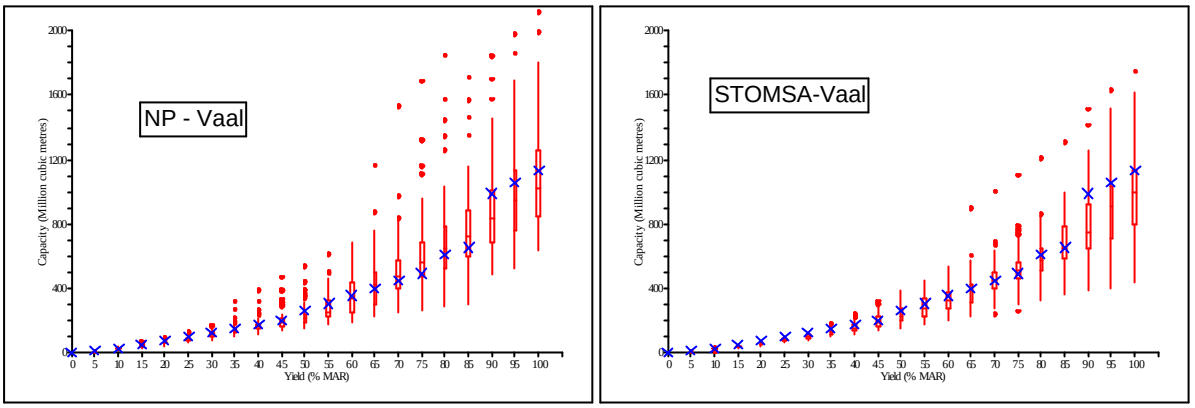


Figure 11 Box plots of capacities for Vaal dam

6. Discussion and Conclusions

The verification tests show that the nonparametric (NP) model reproduces the statistics of the historical sequences consistently and also provides a considerable range of the statistics as probabilistic analysis of water resource systems requires. The five objectives listed in the introduction were therefore accomplished. The validation tests show the NP method to be satisfactory although has a tendency to lead to capacity overestimates at high yields with some sites.

The **STO**chastic **M**odel of **S**outh **A**frica (STOMSA) was found to be robust especially in the yield - storage capacity test. STOMSA however has the limitation of time series methods mentioned by Sharma and O'Neill (2002) of not maintaining the monthly (or seasonal) serial correlations between the end of one year and the beginning of the next. This limitation was observed with all the five sequences tested here while the NP model was found to reproduce these correlations effectively. For situations where reproducing this correlation is important, NP may therefore be preferred to STOMSA. It may be possible that STOMSA could deal with the problem by applying the dissagregation method used for NP or a similar approach. Srinivas and Srinivasan (2001) found a hybrid that borrows the good features of parametric and nonparametric methods to perform better. It was also found that STOMSA had the tendency to underestimate the annual serial correlation while NP did not show any bias.

As a way forward, it is recommended that the NP model be tested on larger problems than used in this work and the comparison with STOMSA be continued. It would be appropriate to include an assessment of varying the ranges of the five parameters of the NP model in any future work. Further improvements, if need be, and recommendations regarding its value for practical application can then be made.

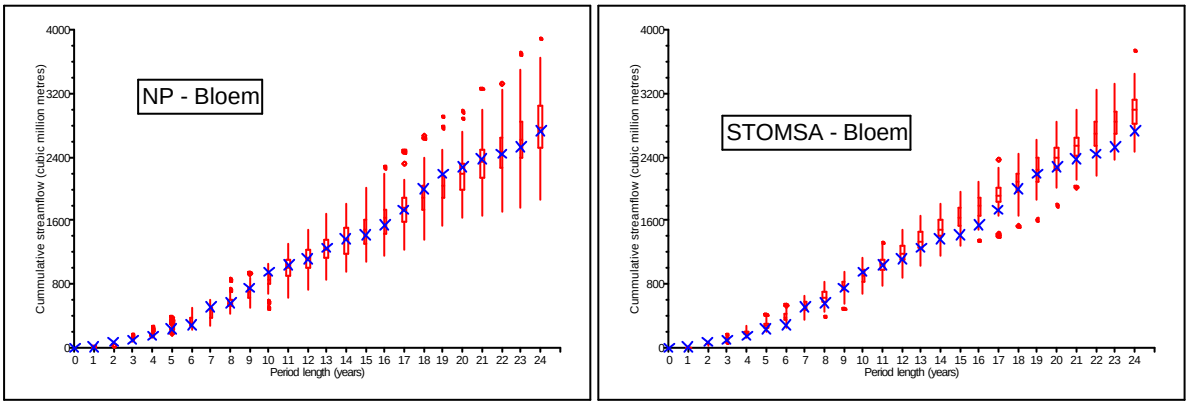


Figure 12 Box plots of minimum cummulative streamflows for Bloemhof dam

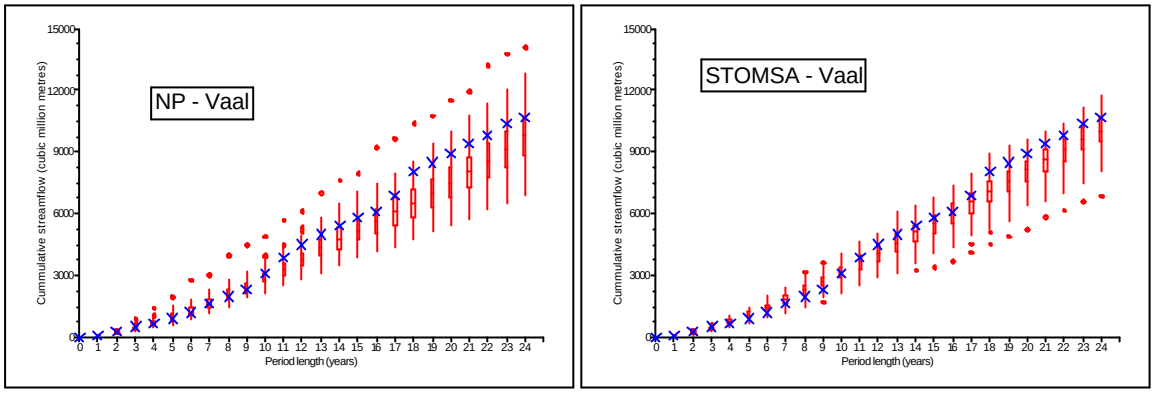


Figure 13 Box plots of minimum cummulative streamflows for Vaal dam

References

Basson MS, Allen RB, Pegram GGS and Rooyen JA, Probabilistic management of water resource and hydropower systems, Water Resources Publications, 1994, 424pp.

Lall U and Sharma A, 1996, A nearest neighbor bootstrap for resampling hydrologic time series, Water Resour. Res., 32(3), 679-693.

Pegram GGS and McKenzie RS, January 1991, Synthetic streamflow generation in the Vaal River System Study, The Civil Engineer in South Africa, 15-24.

Sharma A, Tarboton DG and Lall U, 1996, Streamflow simulation: A nonparametric approach, Water Resour. Res., 33(2), 291-308.

Sharma A and O'Neill R, 2002, A nonparametric approach for representing interannual dependence in monthly streamflow sequences, Water Resour. Res., 38(7), 10.1029/2001WR000953, 5-1 – 5-10.

Srinivas VV and Srinivasan K, 2000, Post-blackening approach for modelling dependent annual streamflows, J. Hydrol., 230(1-2), 86-126.

Srinivas VV and Srinivasan K, 2001, A hybrid stochastic model for multiseason streamflow simulation, Water Resour. Res., 37(10), 2537-2549.

Tarboton DG, Sharma A and Lall U, 1998, Dissagregation procedures for stochastic hydrology based on nonparametric density estimation, Water Resour. Res., 34(1), 107-119.

Van Rooyen P and McKenzie R, 2004, Monthly Multi-Site Stochastic Streamflow Model, STOMSA User Guide, WRC Report No. 909/1/04.

Vogel RM and Shallcross AL, 1996, A moving blocks bootstrap versus parametric time series models, Water Resour. Res., 32(6) 1875-1882.